International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

# Smart Crop Recommendation System with Plant Disease Identification

***Chaytanya Datey[1], Dharmendra Shukla[2], Harsh kumar[3], Bikesh Yadav[4], Prof. Madhu Yadav[5]***

[1] Dept. of CSE SSTC, Bhilai, India
[2] Dept. of CSE SSTC, Bhilai, India
[3] Dept. of CSE SSTC, Bhilai, India
[4] Dept. of CSE SSTC, Bhilai, India
[5] Dept. of CSE SSTC, Bhilai, India

**ABSTRACT—**

Agriculture remains the backbone of many economies, including India, providing livelihoods to millions. However, it faces persistent challenges such as climate change, soil degradation, and plant disease outbreaks. This study presents the development of a web-based application that offers real-time crop recommendations using machine learning models. The system analyzes key environmental parameters—soil nutrients, temperature, humidity, pH levels, and rainfall—to support optimal crop selection. Seven machine learning models—Decision Tree, Naive Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, XGBoost, and K-Nearest Neighbors (KNN)—were trained and evaluated. Among these, the Random Forest model demonstrated the highest accuracy, making it the most effective for crop forecasting. Additionally, the application integrates a Plant Disease Identification module using a Convolutional Neural Network (CNN), which classifies diseases based on leaf image analysis, enabling early detection and intervention. This research aims to empower farmers with accessible, AI-driven tools for data-informed agricultural decisions, ultimately enhancing productivity and sustainability.

**Keywords**: Agriculture, Crop Recommendation, Machine Learning, Random Forest, Convolutional Neural Network (CNN), Plant Disease Detection, Precision Farming, Artificial Intelligence, Climate Resilience, Smart Farmingx

## I. INTRODUCTION

The integration of machine learning (ML) and artificial intelligence (AI) technologies has ushered in a new era across various sectors, with agriculture emerging as a key beneficiary. Among the diverse applications of these technologies in the agricultural domain, smart crop recommendation and plant disease identification systems have proven to be transformative. These advancements are enabling data-driven decision-making that enhances productivity, sustainability, and efficiency in farming practices.

The use of ML and data science has revolutionized traditional agricultural methods. One such innovation is the development of crop recommendation systems that leverage predictive models to provide tailored guidance to farmers regarding crop selection. These systems analyze various environmental and soil-related parameters—such as nutrient levels, rainfall, pH, and location—to suggest the most suitable crops for a given region. In addition, modern image classification techniques, particularly Convolutional Neural Networks (CNNs), have enabled accurate identification of plant diseases through leaf image analysis, allowing for timely intervention and disease management.

In this paper, we present a comprehensive Crop Recommendation and Plant Disease Identification System implemented using Python. The system incorporates multiple machine learning algorithms, including Decision Tree, Naive Bayes, SVM, Logistic Regression, Random Forest, XGBoost, and K-Nearest Neighbors (KNN), for classification tasks. Among these, the Random Forest model exhibited the highest accuracy in crop prediction tasks. For disease detection, a CNN-based model was developed to analyze leaf images and classify common plant diseases.

The system is deployed via a user-friendly web interface built with Streamlit, allowing users to easily input essential data such as nitrogen, phosphorus, potassium levels, rainfall, pH, and temperature, and receive crop recommendations. It also supports the upload of plant leaf images for disease diagnosis. This integrated solution is designed to support farmers and agricultural stakeholders in making informed decisions that enhance crop yield and reduce losses due to disease.

## II. LITERATURE REVIEW

The emergence of smart farming technologies has revolutionized the agricultural landscape by introducing intelligent solutions for enhancing crop productivity, optimizing resource utilization, and mitigating crop diseases. Among these technologies, crop recommendation systems and plant disease identification tools have gained significant attention for their ability to provide data-driven insights and enable timely interventions. Early disease detection, in particular, plays a crucial role in preventing the spread of infections and reducing crop losses, contributing to more resilient and sustainable farming practices .

### *A. Machine Learning Algorithms in Agriculture*

The integration of machine learning (ML) into agricultural systems has opened new avenues for intelligent crop prediction, management, and disease control. Various algorithms have been applied and evaluated for their effectiveness in supporting informed decision-making in agriculture . Below, we summarize key ML models that are widely cited in literature for crop recommendation tasks.

**1) Decision Trees**

Decision Trees are widely recognized for their simplicity and interpretability in both classification and regression tasks . In the context of agriculture, they have been used to model relationships between soil and climatic parameters (such as pH, moisture, and temperature) and crop suitability. The hierarchical structure of decision trees makes it easy to visualize decision paths, which enhances their applicability among end-users such as farmers and agronomists. Their effectiveness in capturing nonlinear patterns in agricultural data has been demonstrated across various crop yield prediction studies.

**2) Random Forests**

Random Forests extend the capabilities of Decision Trees by using an ensemble of multiple trees to improve prediction accuracy and reduce overfitting . They operate by generating multiple decision trees during training and aggregating their outputs for more reliable classification. This method is particularly effective for large, complex agricultural datasets involving multiple environmental and soil parameters. Studies have shown that Random Forests provide high accuracy in crop prediction tasks and are robust to noise and missing data, making them a preferred choice in several smart farming systems.

**3) Convolutional Neural Networks (CNN)**

Convolutional Neural Networks (CNNs), inspired by the structure of the human visual cortex, are a class of deep learning models particularly effective in processing grid-like data such as images. In agriculture, CNNs have emerged as a powerful tool for plant disease identification. By analyzing leaf images, CNNs can automatically learn and extract significant features indicative of various plant diseases, enabling accurate and early diagnosis. Their ability to process complex image data without the need for manual feature engineering makes them well-suited for real-world agricultural applications, especially in field conditions or remote sensing scenarios. Multiple studies have demonstrated the efficacy of CNNs in classifying diseases in crops such as tomato, rice, and maize with high accuracy.

### *B. Optimization of Web-Based Crop Recommendation Systems*

In addition to the development of predictive models, the effectiveness of web-based crop recommendation systems also depends on the optimization of their performance and accessibility. Literature emphasizes several strategies for enhancing the usability and adoption of these platforms among farmers and agricultural stakeholders. Ensuring responsive design is crucial for enabling smooth access across different devices, including smartphones, tablets, and desktop systems—an important factor for users in rural or low-resource settings. Additionally, optimizing loading times contributes to a more efficient and satisfying user experience, especially in regions with limited internet connectivity.

Another important consideration in system design is the integration of external data sources, such as real-time weather forecasts and market trends. Such integrations enrich the decision-making framework, providing users with more holistic and current information. These enhancements contribute significantly to the practical deployment and adoption of crop recommendation systems in real-world agricultural settings.

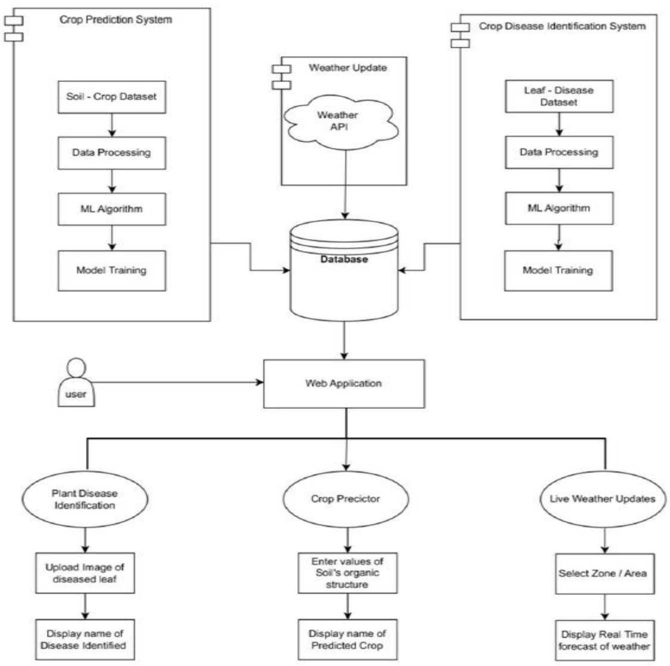



**Fig. 1. System architecture**

## III. METHODOLOGY

### *1. Datasets*

Two distinct datasets were utilized in this study: one for crop recommendation and another for plant disease identification.Crop Recommendation Dataset: This dataset contains 2,200 rows, with each row representing environmental and soil conditions, including Nitrogen (N), Phosphorus (P), Potassium (K), Temperature, Humidity, pH, and Rainfall. The target variable, Label, specifies the most suitable crop to grow under these conditions. The goal is to predict the crop label using machine learning algorithms.

Plant Disease Identification Dataset: This dataset consists of 70,295 leaf images, spanning 38 classes, which include 14 distinct plant species and 26 identifiable diseases. The images are standardized to 128x128 pixels to ensure uniformity. The dataset is approximately 5 GB in size. The machine learning model demonstrating the highest accuracy in training is chosen for further evaluation.

### *2. Data Preprocessing*

Effective data preprocessing is crucial for ensuring the quality and usability of datasets in machine learning models.Crop Recommendation Dataset: The numerical features (NPK values, temperature, humidity, pH, and rainfall) are standardized to remove any biases introduced by differences in their scales.Plant Disease Identification Dataset: Raw images may contain noise and inconsistent lighting, which necessitates preprocessing. Common preprocessing steps include image resizing, normalization, and data augmentation techniques such as rotation, flipping, or zooming. These transformations help ensure better feature extraction and improve the model’s performance by mitigating overfitting.

### *3. Train-Test Split*

The dataset is split into training and testing subsets using the train_test_split() function from Scikit-learn.

Crop Recommendation Dataset: 80% of the data (1,760 records) is used for training, and the remaining 20% (440 records) is reserved for testing.Plant Disease Identification Dataset: 80% of the images are used for training, while 20% are set aside for validation and testing to assess the model's ability to generalize to unseen data.This approach ensures that the models are properly trained while also evaluating their performance on data they have not previously encountered.

### *4. Confusion Matrix and Classification Report*

To assess the performance of the machine learning models, we utilized two key metrics from the Scikit-learn library: the Confusion Matrix and the Classification Report.Confusion Matrix: This matrix provides a detailed breakdown of the performance of a classification model, specifically the frequency of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP). It helps in evaluating the model's ability to correctly classify instances of each class.

Precision: Precision measures the ability of the classifier to correctly identify the relevant instanceSpecifically, it computes the ratio of true positives to the total predicted positives (the sum of true positives and false positives) for each class. This metric is particularly important when the cost of false positives is high.

Recall: Recall (also known as Sensitivity or True Positive Rate) indicates the model's ability to correctly identify all relevant instances in the dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives.

F1-Score: The F1-Score represents the harmonic mean of Precision and Recall. It balances the two metrics, giving a single score to evaluate the model’s performance. An F1-Score of 1.0 indicates the best performance, while 0.0 indicates the poorest.

Accuracy: Accuracy is a fundamental evaluation metric that indicates the proportion of correct predictions relative to the total number of predictions made. It is computed

## IV. RESULTS

Seven classification algorithms have been selected, and their respective accuracies have been assessed. For the second task, the preeminent photograph classification convolutional neural network (CNN) architectures were skilled. The ensuing tables present the corresponding outcomes.

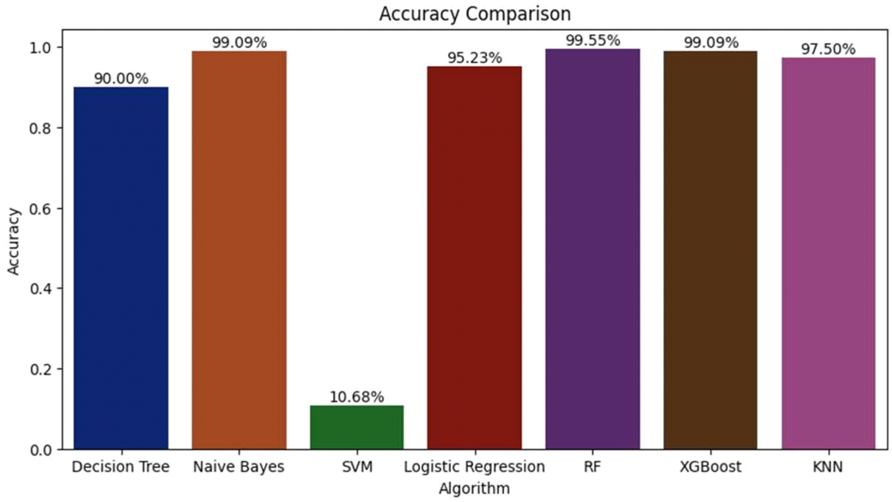



**Fig. 2. Accuracy comparison**

Table 1 offers the accuracy of crop advice responsibilities across diverse algorithms . This table unequivocally demonstrates that the random forest algorithm surpasses all others, accomplishing a remarkable accuracy of 99.54%.

TABLE I. ACCURACY VS ALGORITHMS

| Algorithm | Accuracy |
|---|---|
| Decision Tree | 90.0 |
| Gaussian Naive Bayes | 99.09 |
| Support Vector Machine (SVM) | 10.68 |
| Logistic Regression | 95.23 |
| Random Forest | 99.55 |
| XGBoost | 99.09 |
| KNN | 97.50 |

"Fig. 3" shows that Random Forest models consistently perform well across different crops, often outperforming other models. While some models excel with specific crops, they show varying accuracy overall. Random Forest stands out for its stability and versatility, making it a reliable choice for a wide range of crops. This highlights the model's effectiveness in providing consistent accuracy compared to the more specialized performance of other models.

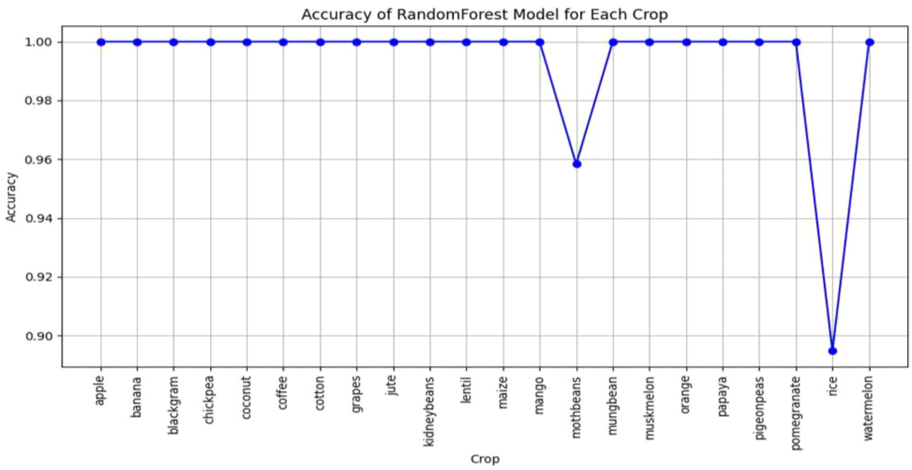



**Fig. 3. Accuracy vs Crop graph for random forest model**

In this study, we employed a convolutional neural network to detect plant diseases from a dataset containing images of various plant species. The dataset, consisting of 17,572 images across 38 distinct classes, was organized into a validation set using the tf.keras.utils.image_dataset_from_directory function, ensuring that each image was resized to 128x128 pixels to match the input requirements of our model. The class names, representing various plant diseases and healthy conditions, were automatically inferred from the directory structure. We utilized a pre-trained CNN model, which was loaded from a saved.keras file, to perform the predictions . Here we use a dropout layer to prevent overfitting by randomly setting a fraction of input units to zero at each update during training.For visualization and testing purposes, a sample image of a potato plant affected by early blight was processed. This image was read, converted from BGR to RGB using OpenCV, and resized to the required dimensions. The preprocessed image was then fed into the model to obtain predictions. The model's output was a probability distribution across all classes, from which the class with the highest probability was identified using the np.argmax function. The predicted class label, in this case, was displayed alongside the original image, confirming the model's ability to accurately identify the plant disease.

This approach demonstrates the efficacy of CNNs in the automated detection of plant diseases, which can significantly aid in early diagnosis and management, ultimately contributing to better agricultural practices. “Fig.4” illustrates an example of a diseased plant image used for plant disease identification.

The detailed features in the image, such as spots, discoloration, and other symptoms, enable the models to learn and identify various plant diseases, contributing to smarter crop management and healthier yields.

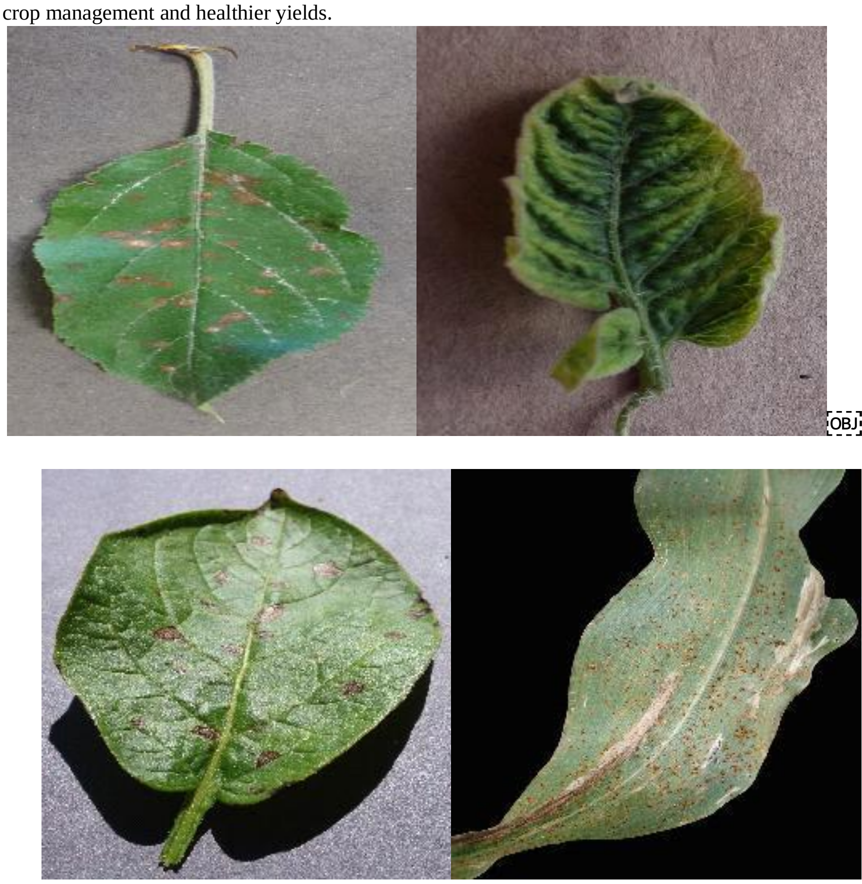
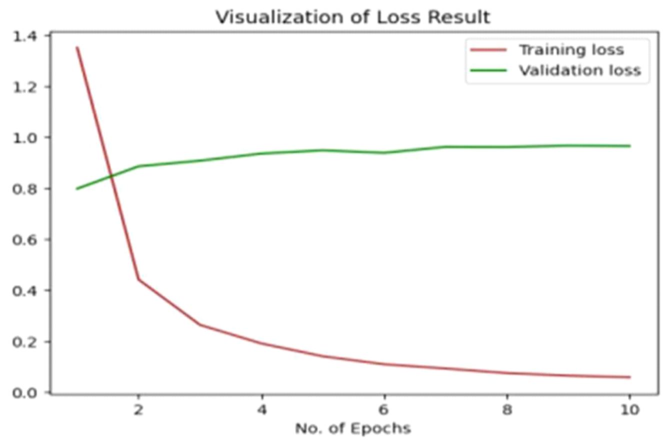


**Fig. 4. Example of diseased image**



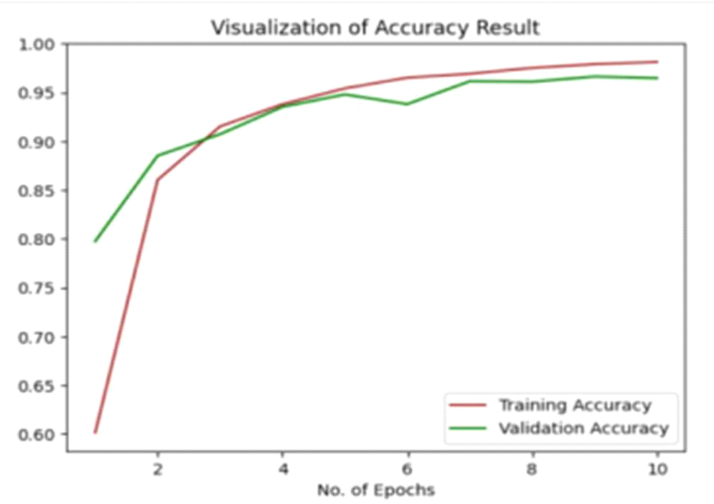



**Fig. 5. Training and Validation (accuracy and loss) simple CNN**

The graph depicts the performance of a plant disease detection model using a convolutional neural network (CNN). The model's training and validation accuracy are plotted alongside the training and validation loss. It is important for the training and validation accuracy to be high, and the training and validation loss to be low . High training and validation accuracy indicate that the model can accurately distinguish between healthy and diseased plants . Low training and validation loss signifies that the model is learning from the training data and generalizes well to unseen data. Learning curves for a deep learning algorithm show how well it is learning the dataset throughout training in an incremental manner.

The accuracy and loss of CNN during training and validation are shown in "Fig.5". Although there was little volatility during the validation test, the loss curve shows that the training and validation losses dropped over time and that the interval between them was short over the experiments. The training and validation accuracy and loss for the proposal are displayed in "Fig.5". In contrast, the loss graph shows strong fitting for both training and validating loss curves .

## VI. CONCLUSION

In conclusion, the project has successfully developed a sophisticated system for smart crop recommendation and disease detection using advanced technologies and techniques. By combining machine learning algorithms and real-time data processing, the project provides insights that can be applied to improving crop management and reducing the impact of plant diseases.

Modular and scalable design patterns adopted while adhering to industry standards ensured the reliability, scalability, and maintainability of the solution. Using best practices of database design and user-centric interface design principles, the project provided farmers, regardless of their technical skills, with a solution that facilitated seamless communication and usability.Overall, precision agriculture has improved dramatically through the integration of strategies, innovation, and adherence to industry standards. The integration of cutting-edge technologies for smart crop recommendations and disease detection has provided farmers with the necessary tools for informed decision-making and sustainable agricultural practices.

n future, the dataset can be frequently updated with new samples and cases. The performance of Crop Recommendation and Plant Disease Identification, implemented by machine learning classification and prediction models, can be improved by integrating it with practices such as multispectral and hyperspectral imaging data for insights on plant health, and Reinforced Learning to optimize the performance of existing strategies. Including the temperature and rainfall data of the region in which the farm is located can also help in suggesting the best-suited crop, potentially reducing crop losses and improving yields. More features, like profit-oriented crop prediction for farmers based on the selling price of the crops and investment needed for that crop, can be added. We can also include crop growth schedule managers to help new farmers get acquainted with the timeline of different processes in farming and to help with watering, fertilizing, and pest cleaning timely and regularly. The concepts of edge computing can also be implemented. By integrating the system with a movable device with a camera and sensors, farmers can also help with real-time monitoring of their farms. Furthermore, the development of a collaborative platform, inviting various farmers, researchers, and agronomists to share data and insights on various cases and info about any new ones, will also have a significant impact. Such crowd-sourced data collection could also help in significantly improving model accuracy.

### VII. REFERENCES

[1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in Proc. ACM CSCW, 1994, pp. 175–186.

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513–523, 1988.

[3] J. Bennett and S. Lanning, "The Netflix Prize," in Proc. KDD Cup and Workshop, 2007.

[4] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, vol. 12, no. 4, pp. 331–370, 2002.

[5] B. Sarwar et al., “Item-based collaborative filtering recommendation algorithms,” in Proc. 10th Int. Conf. World Wide Web, 2001, pp. 285–295.

[6] P. Covington, J. Adams, and E. Sargin, “Deep Neural Networks for YouTube Recommendations,” in Proc. ACM RecSys, 2016, pp. 191–198.

[7] G. Shani and A. Gunawardana, “Evaluating recommendation systems,” in Recommender Systems Handbook, F. Ricci et al., Eds., Springer, 2011, pp. 257–297.

[8] GroupLens Research, “MovieLens 32M Dataset,” May 2024. [Online]. Available: https://grouplens.org/datasets/movielens/

[9] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[10] W. McKinney, “Data Structures for Statistical Computing in Python,” in Proc. 9th Python in Science Conf., 2010, pp. 51–56.

[11] P. Grinberg, Flask Web Development: Developing Web Applications with Python, O’Reilly Media, 2018.